IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 4, April 2017

Feature Extraction Mining Method For Intrusion Detection Systems based on Darpa 1999 Dataset

Karim Hashim Al-Saedi[1], Nazhat SaeedAbdulrazzaq[2], Dhiya Ibraheem Selman[3]

University of Mustansiriyah, Baghdad, Iraq[1]

University of Technology, Baghdad, Iraq[2]

Informatics Institute for Postgraduate Studies,UITC, Baghdad, Iraq[3]

Abstract: Intrusion Detection Systems (IDSs) have turned into an important security system for managing dangers and a needful aspect of global security architecture.The IDS is triggering alerts for any suspicious activity which means thousand alerts that the analysts should take care of it. These Alerts has contained irrelevant and redundant features and most of them don't require big attention by researches. Deleting the alert attributes or reducing the amount of them from the entire amount alert attributes lead the researchers to create many methods such as principle component analysis. Feature ExtractionMining Method in IDS is an important data mining step.In this paper, we focus on an approach of feature selection based on Darpa 1999. The first step that based on data preprocessing and configuration for the next stage and guides the initialization of search process for the second step that based on principle component analysis whose outputs the final feature subset.

Keywords: Intrusion Detection system(IDS), Feature Extraction, Data Mining(DM), Principle Component Analysis.

I. INTRODUCTION

Daily Internet and web applications sneak through businesses and life. In today highly competitive market, everyone wants to utilize Internet for their needs. Corporate uses Internet for growing the business by cooperation and communication. An individual utilizes Internet for social and personal objectives. As a result of the wide benefits of the Internet, for this reason, critical security issues to our doorstep. Computers networks become part of business network in today's communication age. Most modern businesses cannot continue or at least cannot advance effectively without Internet [1]. On the other side, During the last decade with the increasing of cyber attacks, data safety has become an important and first issue all over the world [2]. An intrusion detection system is a security technique that monitors and analyzes network packets to provide real-time warnings to unauthorized access to system resources or to archive log and traffic information for later analysis [3]. Newly, several security systems and tools are utilized in networks to provide security such as Intrusion Detection Systems (IDS). An IDS is utilized to discover all intrusions in an efficient manner and when it notice any suspicious event representing a threat or abnormal behavior which may outcome in damaging computer networks and systems it generates alerts**[4].**Intrusion detection system trigger a great volume of alerts by discovering these intrusions and causes problems during the analysis process. Each alert is consists of set of attributes:sensor, alert type, classification, priority, date, time (hours ,minutes, seconds and milliseconds), source IP address, destination IP address, source port number, destination port number, protocol, TTL, TOS, ID, Iplen, Dgmlen, type, code and packet type. The reduction of these attributes has become a necessary condition for many researchers [5]. Many Researches [6] in false positive alerts reduction process depending on some attributes without using feature selection or extraction methods. They based on some of the important attributes that depend on it e.g., priority, Ip addresses for each source and destination, source and destination ports and protocol alert type. Feature selection is a pre-processing data mining technique that discovers a minimum subset of features that captures the relevant properties of a dataset. these techniques are very useful for increasing and improving the performance oflearning algorithms, feature selection has been widely utilized [7].

The remaining part of this paper organized as follows: next section presents problem statement, and then paper presents Methodology of Feature extraction. Finally, a discussion about theresulted feature extraction method and the obtained results presented followed by conclusions.

II. PROBLEM STATEMENT

Intrusion detection system IDS deals with large amount of data, which contains various irrelevant and redundant features. These alert attributes causing slow training and testing process, higher resource consumption and increasing computational time as well as attributes select is important step to help in false positive alerts reduction rate.

DOI10.17148/IJARCCE.2017.64147

III.METHODOLOGY OF FEATURE EXTRACTION MINING

Feature Extractionis the process of choosing a subset of original features so that the feature space is optimally reduced to evaluation criterion. In this section, we explain in detail the seven key steps as shown in Fig. 1.

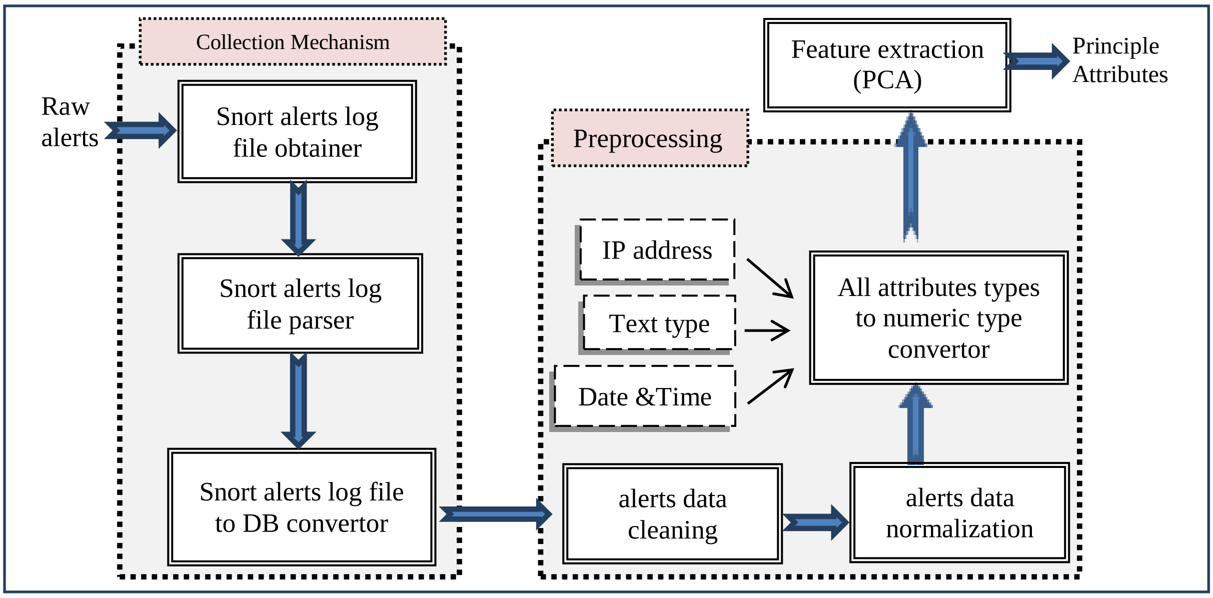
Collection Mechanism
Raw alerts
Snort alerts log file obtainer
Snort alerts log file parser
Snort alerts log file to DB convertor
Feature extraction (PCA)
Principle Attributes
Preprocessing
IP address
Text type
Date &Time
All attributes types to numeric type convertor
alerts data cleaning
alerts data normalization

Fig.1 The main steps of feature selection mining

A. Collection Mechanism

The goal of this step is to Preparing and configuring data to be ready for processing and use to the next stage, as well as to determine the several data formats and share data of interest with the next stage (preprocessing). Collection Mechanism has three main components: Obtain Snort alerts log file, Snort Alerts Log File Parser and Convert Snort Alerts Log File To database as shown in fig. 2.

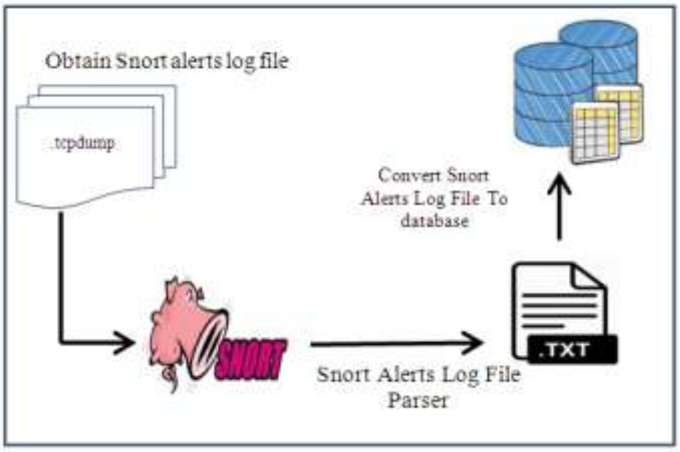
Obtain Snort alerts log file
tcpdump
SNORT
Snort Alerts Log File Parser
.TXT
Convert Snort Alerts Log File To database

Fig. 2 Collection Mechanism

A.1 Obtain Snort Alerts Log File

In this work, the raw log file was collected from download and install snort software by snort website and using standard dataset called training Darpa 1999 that been in tcpdump format. The file written common txt log format and consists of 21 attributes (in addition to alert_id) such as sensor, alert type (type), classification (classf), priority(prio), date, time (hours ,minutes, seconds and milliseconds), source IP address(ip_s), destination IP address(ip_d), source port number (port_s), destination port number(port_d), protocol, TTL, TOS, ID, Iplen, Dgmlen, type (type1), code and packet type(endtype) as shown in fig. 3 [8]. For the analysis purposes, the data retrieved from this file needed to be preprocessed.

```
[**] [1:384:5] ICMP PING [**]
[Classification: Misc activity] [Priority: 3]
06/15-14:58:16.883844 192.168.1.5 -> 192.168.1.1
ICMP  TTL:32  TOS:0x0  ID:2711  IpLen:20  DgmLen:50
Type:8  Code:0  ID:8   Seq:24841  ECHO
```

Fig. 3Sample of full mode alert Format

A.2 Snort Alerts Log File Parser

In this block, the separator characters and format of log file must be determined indeed, there are many characters are used as a delimiter (split character) in log file such as space, dot,colon and others. Therefore, this characters should be determined to the method.

A.3 Convert Snort Alerts Log File To database

After reading the Snort log file, the Snort log file data will be transferred to SQL Server relational database in order to make it appropriate to apply the feature extraction technique in the next stage of the process. One of the main problems encountered when dealing with the log file and converting it to database are amount. Algorithm for converting alerts log file to table in database is shown in fig. 4.

```
Input : Alerts Log File
Output: Log Table (LT)
Begin
1)  Open a DB connection
2)  Create a table to store log file
3)  Open log file
4)  While not end of log file
        a)   Read an entry of log file
        b)   Tokenize the fields depending on delimiter character
        c)   Insert all fields into the Log Table (LT)
5)   End while
6)   Close a DB connection and Log File .
End.
```

Fig. 4 Alerts log file to DB convertor algorithm

B. PREPROCESSING

The second stage of system is preprocessing. After Preparing Data from the previous stage, data is become ready for processing. This stage is responsible for extracting the standard features from the IDS alert file after the first component has performed its function. This consists of three subcomponents, namely, Alerts Data Cleaning, Convert All Attributes Types To Numeric Type,Alerts Normalization.

B.1 Alerts Data Cleaning

```
Input : Alerts Log File without ICMP port numbers
Output: Log Table (LT) with -1
Begin
1)  Open log file
2)  While not end of log file
    a)  Read an entry of log file
    b)  While line != " "
    c)  Read lines until line3
    d)  If number x before  (->) symbol even (:)symbol then Source Port number = x
           Else Source Port number = -1
    e)  If number x between last (:)symbol and end line3 then  destination Port number = x
           Else destination Port number = -1
    f)  End whiles
End
```

Fig. 5 Alerts log file cleaner Algorithm

It is the first stage from data preprocessing. There are many service protocols e.g. ICMP, TCP and UDP in dataset. Because ICMP is used, port attributes for each source and destination in alerts is not include, so they replaced by -1 value because It does not exist between port ranges (0 - 65536) as shown in figure (5).

B.1 Convert All Attributes Types To Numeric Type

In fig. 6, Alert attributes are in the form of numerical and non-numerical values. Attributes that contain numerical values are Alert_id, Sensor, SourcePort, DestinationPort, and DateTime.The rest are non-numerical values (i.e., SourceIPaddress,DestinationIPaddress, ServiceProtocol and lertType ..etc) and have to be mapped into numerical values.For instance to convert a 32-bit IP address (IPaddr) which in $X_1.X_2.X_3.X_4$ format, mapping as (1) was used.

$$IPaddr = ((X_1 * 256 + X_2) * 256 + X_3) * 256 + X_4 \quad (1)$$

Preprocessing unit converts string values of attributes of alert to numerical data as shown in(2)

$$\text{ServiceProtocol} = \begin{cases} 1\text{, protocol} = \text{ICMP} \\ 2\text{, protocol} = \text{TCP} \\ 3\text{, protocol} = \text{UDP} \end{cases} \quad (2)$$

```
Input : Log Table (LT) with multiple attributes types
Output: Log Table (LT) with numeric attributes types
Begin
1.  Open log table
2.  Read all attributes (p)
3.  For i =1 to p
4.  If attribute(i) is numeric then   No thing;
5.  if attribute(i) is text then Replace each unique value in
    attribute(i) by unique number ; repeat ;
6.  if attribute(i) is date/time then using datenum function;
7.  If attribute(i) is '$X_1.X_2.X_3.X_4$' then
8.  IPaddr = (($X_1$ x 256 + $X_2$) x 256 + $X_3$) x 256 + $X_4$;
9.  End if
End
```

Fig. 6 Convert All Attributes Types to Numeric Type Algorithm

B.3 Alerts Normalization

It is used for Normalizing the data attempts to give all attributes an equal weight.In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean μ (average) and standard deviation σ of A. A value, v_{old}, of A is normalized to v_i by computing as in equations (3), (4) and (5) [9]:

$$\mu = \sum_{i=1}^{n} Xi/n \quad (3)$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N} \frac{(Xi-\mu)}{N-1}}$$

$$V_{new} = (V_{old} - \mu) / \sigma \quad (4)$$

Alert normalization is explained in fig. 7.

```
Input : Log Table (LT)
Output: Log Table (LT)
Begin
1.  Open log table
2.  Read all attributes (p)
3.  For i =1 to p
4.  count mean μ for each column in dataset by using Eq.(3).
5. count standard deviation σ for each column in dataset by using Eq.(4).
6. apply equation (4) for each value in column by using Eq.(5).
7. repeat steps (4), (5) and (5)  until the final column in dataset.
8. End For
End
```

Fig. 7 Z-Score Normalization Algorithm

DOI10.17148/IJARCCE.2017.64147

c. Principal Components Analysis (PCA)

It is a statistical analysis method that converts multiple variables into fewer main variables. It generally is utilized to data compression, regression analysis, system evaluation and weighted analysis so on [10]. Here are eight steps to get the PCA for a given data [11][12][13].

```
Input : Log Table (LT) with 20 attributes
Output : Log Table (LT) with 13 attributes
Begin
1) Get log table (LT) from the previous stage.
2) Read all attributes (p) except alert_id.
3) Calculate number of rows (n) and columns (p).
4) Center the data by using the following equation: $V_{new} = (V_{old} - \mu)$.
5) Find covariance matrix (p*p).
  a) Create ones vector (one_vector) by size (1,n)
  b) Calculate mean for each column by using Eq.(3)
  c) Subtract LT from mean in each column:
  d) run covariance equation: $\sum(X - \mu_x)( Y - \mu_y ) /n-1$
6) Find eigenvalues and eigenvalues : $(A. x = x. \lambda)$
  a) Extract the main diagonal elements of eigval:$\lambda_1,\lambda_2,\ldots,\lambda_p$
  b) Order $\lambda$ in descending order: $\lambda_1 < \lambda_2 < \ldots < \lambda_p$
  c) In reverse, Order eigenvectors.
7) Plot scree graph, to decide how many components to retain.
8) Select values greater than 0.5 (columns names)
End.
```

Fig. 8 Principal Components Analysis algorithm

As shown in fig. 8, PCA can be utilized to compress data sets of high dimensional vectors into lower dimensional ones. It displays a smaller-dimensional linear representation of data vectors such that the original data could be reconstructed from the compressed representation [14].Objective of PCA is detected or decreased the dimensionality of the data set and to define new meaningful underlying variables. The mathematical technique utilized in PCA is called Eigen analysis: by solve for the eigenvalues and eigenvectors of a square symmetric matrix (covariance matrix) with sums of squares and cross products [15].

IV. EXPERIMENTAL REDUCTION RESULTS AND ANALYSIS

The proposed approach was implemented on Microsoft windows 7 64 bits service pack 1 and AMD dual core i5 2.5 GHz of CPU and 4 GB of RAM.

We implemented the proposed framework by leveraging Matlab version R2013a with Microsoft Excel 2010. The next section explains the data sets utilized in our system.

A. Darpa 1999 Dataset

Intrusion Detection Systems (IDS) dataset is the DARPA/Lincoln Laboratory off-line evaluation data set. Since the two evaluations in 1998 and 1999. It is available in website https://www.ll.mit.edu/ideval/data/1999data.html .

B. Accomplishment experiments

The data sets in our check leverage IDS Snort 2.9 . To generate an acceptable alert set, we used IDS Snort 2.9 in our experiments, which has the flexibility of providing alerts in a flat file. The experiment was done on the 1226 alerts are selected from training Darpa 1999 dataset. after applied PCA algorithm, The important step (as shown in figure(8)) is covariance matrix S that is a Symmetrical matrix by 20*20 size and the main diagonal is variance each attribute in dataset.

The next step is find eigenvalues and eigenvectors. The eigenvalues is 20*20 matrix that all its elements equal zero except the main diagonal elements (eigenvalues). The main diagonal elements is concluded then rearrange order in descending order.

As shown in Figure (9), The first three eigenvalues form a steep curve followed by a bend and then a straight-line trend with shallow slope. The recommendation is to retain those eigenvalues in the steep curve before the first one on the straight line.

Thus in Figure 4.8, three components would be retained.In table (1), The first three principal components accounted by Scree graph as shown in figure (9) . The corresponding eigenvectors are as follows:

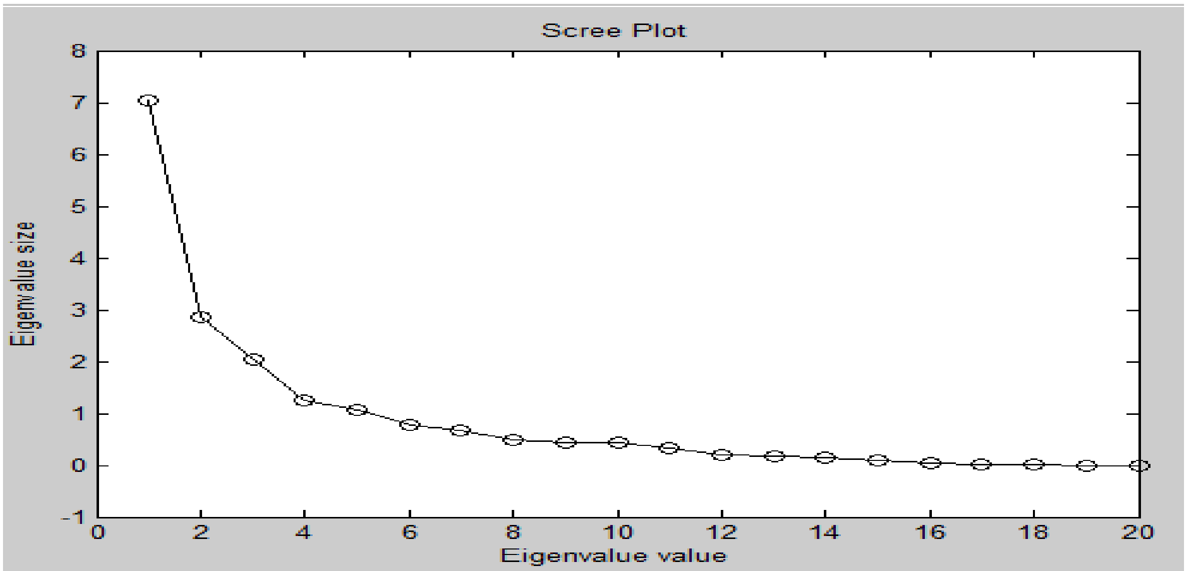

Fig. 9 Scree graph for eigenvalues

TABLE 1 THE CORRESPONDING EIGENVECTORS FOR SCREE GRAPH

attributes	Component-1	Component-2	Component-3
Sensor	-0.3027	0.1048	-0.1548
Type	0.2866	0.1078	-0.2758
Classf	-0.2028	0.0224	0.4664
Prio	0.3757	-0.0635	-0.0272
Date	0.2095	0.0531	0.2021
Hm	-0.0509	-0.1478	-0.1235
Spart	-0.1570	0.0660	0.3080
Ip_s	0.1978	-0.0388	0.4638
Port_s	-0.3188	0.0567	0.0542
Ip_d	0.0374	-0.0267	0.5390
Port_d	-0.3280	0.0724	0.0319
Protocol	0.3645	-0.0682	0.0951
Ttl	0.1137	0.5060	-0.0168
Tos	4.0658e-20	0	0
Id	-0.3202	0.0538	-0.0716
Iplen	1.7347e-18	3.4694e-18	1.1102e-16
Dgmlen	-0.2466	0.0605	0.0022
Type1	-0.0511	-0.5773	-0.0333
Code	-0.0828	-0.0333	0.0707
endtype	0.0671	0.5758	0.0232

We will take the absolute values in table (1) because the negative sign meaning the direction of the movement. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component.

The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component.Because that values in table (1) is a little, we will multiply weights for each component in table (1). The weighted vector is [1.6 ; 1 ; 2.7] and table (1) will be as shown in table (2):

DOI10.17148/IJARCCE.2017.64147

TABLE 2 THE FINAL RESULT OF PCA AFTER MULTIPLYING THEM BY WEIGHTED VECTOR

attributes	Component-1	Component-2	Component-3
Sensor	0.4843	0.1048	0.2864
Type	0.4586	0.1078	0.5102
Classf	0.3244	0.0224	0.8628
Prio	0.6011	0.0635	0.0503
Date	0.3352	0.0531	0.3739
Hm	0.0815	0.1478	0.2284
Spart	0.2512	0.0660	0.5699
Ip_s	0.3165	0.0388	0.8581
Port_s	0.5102	0.0567	0.1003
Ip_d	0.0598	0.0267	0.9972
Port_d	0.5248	0.0724	0.0590
Protocol	0.5831	0.0682	0.1760
Ttl	0.1819	0.5060	0.0311
Tos	6.5052e-20	0	0
Id	0.5124	0.0538	0.1324
Iplen	2.7756e-18	3.4694e-18	2.0539e-16
Dgmlen	0.3945	0.0605	0.0042
Type1	0.0818	0.5773	0.0617
Code	0.1325	0.0333	0.1307
endtype	0.1073	0.5758	0.0428

Through the results that displayed in table (2). the variables that have up of 0.5value will use in the next stage as shown in table (3). This value selected because mean in Gaussian normal is 0.5 that be between 0 and 1.

TABLE 3 THE FINAL RESULTS FOR PCA BASED ON UP 0.5 VALUES

attributes	Component-1	Component-2	Component-3
Type	0.4586	0.1078	0.5102
Classf	0.3244	0.0224	0.8628
Prio	0.6011	0.0635	0.0503
Spart	0.2512	0.0660	0.5699
Ip_s	0.3165	0.0388	0.8581
Port_s	0.5102	0.0567	0.1003
Ip_d	0.0598	0.0267	0.9972
Port_d	0.5248	0.0724	0.0590
Protocol	0.5831	0.0682	0.1760
Ttl	0.1819	0.5060	0.0311
Id	0.5124	0.0538	0.1324
Type1	0.0818	0.5773	0.0617
endtype	0.1073	0.5758	0.0428

V. CONCLUSION AND RESULTS

Through the results that displayed in table (3). We noticed that PCA can reduce the number of attributes in dataset attributes). After using PCA method , the number of attributes that displaying are 13 attributes. This meaning that the PCA can remove irrelevant attributes by elimination rate almost 35% with retain the main and important attributes. The final results are alert Type, classification, priority, millisecond , source IP address (ip_s), destination IP address(ip_d), source port number (port_s), destination port number (port_d), protocol, TTL, Id, type, packet type (endtype) attributes. Elimination (Reduction) Rate is shown in fig.10.

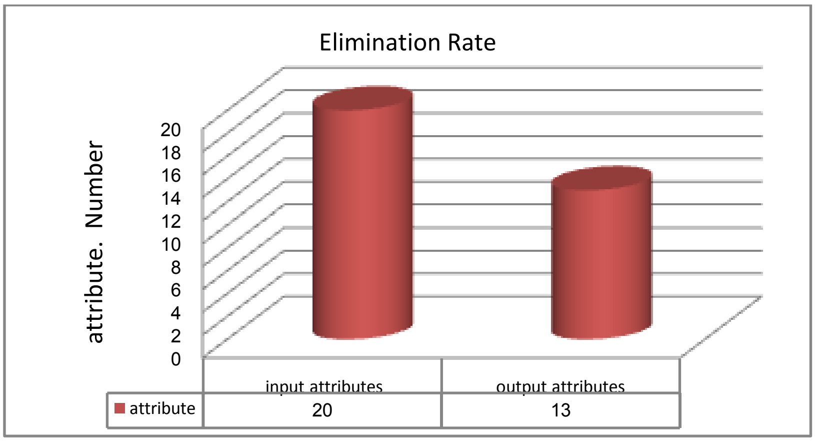

Fig. 10 Elimination Rate

REFERENCES

[1] Dharmendra G. Bhatti and P. V. Virparia ,"Data Preprocessing for Reducing False Positive Rate in Intrusion Detection ", International Journal of Computer applications,Vol, 57, No.5, November 2012

[2] AsiehMokarian, Ahmad Faraahi, ArashGhorbanniaDelavar," False Positives Reduction Techniques in Intrusion Detection Systems-A Review", International Journal of Computer Science and Network Security, vol.13 No.10, October 2013

[3] Dr.Soukaena H. Hashim and Inas A. Abdulmunem," A Proposal to Detect Computer Worms (Malicious Codes) Using Data Mining Classification Algorithms", Eng. & Tech. Journal, Vol.31, No.2, 2013

[4] El MostaphaChakir and ChancerelCodjovi," False Positives Reduction in Intrusion Detection Systems Using Alert Correlation and Data mining Techniques", International Journal of Advanced Research in Computer Science and Software Engineering, Vol.5, Issue 4, 2015.Available online at: www.ijarcsse.com

[5] KarimAlsaedi, SureswaranRamadass, AmmarAlmomani and SelvakumarManickam,"Collection Mechanism and Reduction of IDS Alert", International Journal of Computer Applications ,Vol 58, No.4, November 2012, https://www.researchgate.net/publication/237065200

[6] Safaa O. Al-Mamory and Hongli Zhang," Intrusion detection alarms reduction using root cause analysis and clustering", Computer Communications, vol.32 , pp.419- 430, 2009.

[7] Zahra Karimi, Mohammad Mansour and Ali Harounabadi, "Feature Ranking in Intrusion Detection Dataset using Combination of Filtering Methods", International Journal of Computer Applications, Vol. 78 , No.4, September 2013.

[8] HomamReda El-Taj and Omar AmerAbouabdalla," False Positive Reduction by Correlating the Intrusion Detection System Alerts: investigation Study ", Journal of Communication and Computer,Vol. 7, No.3, 2010, https://www.researchgate.net/publication/210195144

[9] Jiawei Han, MichelineKamber and Jian Pei," Data Mining Concepts and Techniques ", 3rd Edition, 2012.

[10] Xiaoping Xie, " Principal component analysis-based sports dance development influence factors research", Journal of Chemical and Pharmaceutical Research, 6(7):970-976, 2014. Available online : www.jocpr.com.

[11] Eyad. I. Abbas, " Effect of Eigenfaces Level On The Face Recognition Rate Using Principal Component Analysis ", Eng. &Tech.Journal, Vol.33, Part (A), No.3, 2015

[12] Bruce L. Brown,Suzanne B. Hendrix, Dawson W. Hedges,And Timothy B.Smith,"Multivariate Analysis For The Biobehavioral And Social Sciences", John Wiley & Sons,2012

[13] Liton Chandra Paul, Abdulla Al Suman and Nahid Sultan," Methodological Analysis of Principal Component Analysis (PCA) Method", International Journal of Computational Engineering & Management, Vol. 16, Issue 2, March 2013

[14] Alexander Ilin and TapaniRaiko," Practical Approaches to Principal Component Analysis in the Presence of Missing Values", Journal of Machine Learning Research, vol.11, pp:1957-2000, 2010

[15] Liton Chandra Paul, Abdulla Al Suman and NahidSultan,"Methodological Analysis of Principal Component Analysis (PCA) Method", IJCEM International Journal of Computational Engineering & Management, Vol. 16 Issue 2, March 2013

DOI10.17148/IJARCCE.2017.64147